САНКТ - ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
Математико-механический факультет
Кафедра системного программирования

Тема

Нахождение сайтов начала репликации в ДНК человека

Курсовая работа студента 445 группы
Ромашкина Амира Сергеевича
Научный руководитель,
Доцент АУ, к.м.н, **Юрий Порозов**/

Санкт-Петербург
2011

Оглавление

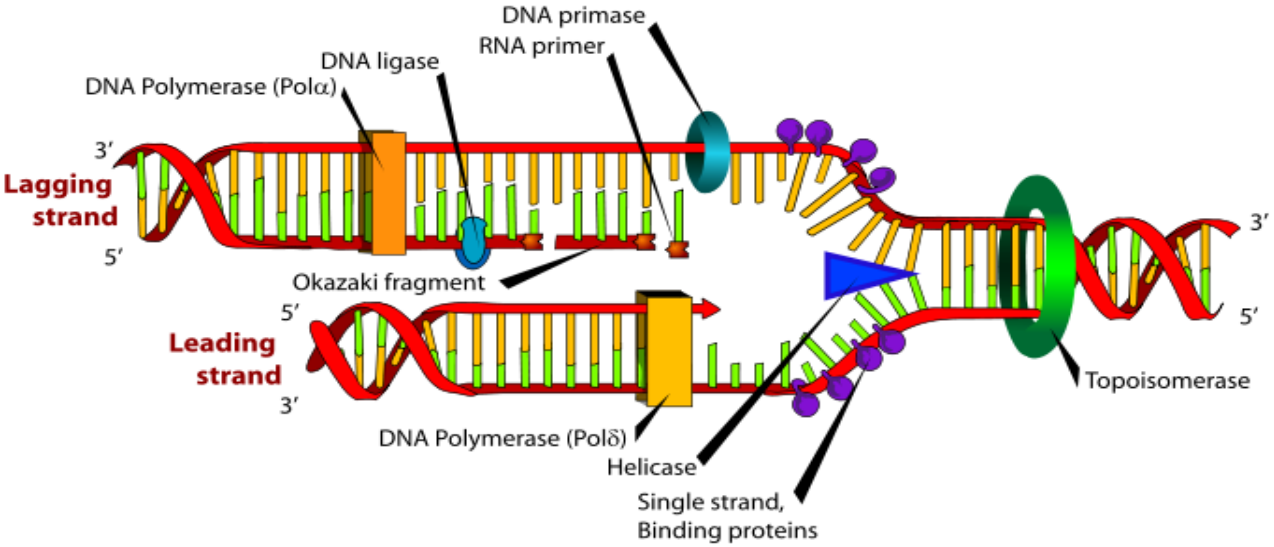
Введение

Биология и информатика долгое время жили отдельно, и биологам самим приходилось выполнять неавтоматизированные сложные экспериментальные расчеты, но это должно было когда-нибудь закончиться. Несколько лет назад биология все-таки тесно слилась с информатикой, образовав новую сферу научной деятельности - биоинформатику. Эта область быстро стала популярной, завоевав свою престижную нишу в техническом мире. Об этом говорят гранты огромных размеров, выдающиеся на исследования в этой сфере и количество программистов, тесно связавших себя с биологией.

Биоинформатика – это довольно новая предметная область и, соответственно, это очень широкое поле для деятельности и фантазии. Мне была предложена одна из интересных задач, касающаяся репликации ДНК.

Обзор предметной области

Репликация ДНК



Репли́кация ДНК — процесс синтеза дочерней молекулы дезоксирибонуклеиновой кислоты, идущий во время синтетической фазы жизненного цикла клетки на матрице родительской молекулы ДНК. При этом генетический материал, зашифрованный в ДНК, удваивается и в процессе последующего деления делится между дочерними клетками. Репликацию ДНК осуществляет сложный ферментный комплекс, состоящий из 15-20 различных белков.

Ферменты и ДНК-связывающие белки расплетают ДНК, удерживают матрицу в разведённом состоянии и вращают молекулу ДНК. Правильность

репликации обеспечивается точным соответствием комплементарных пар оснований и активностью ДНК-полимеразы, способной распознать и исправить ошибку.

Цепи молекулы ДНК расходятся, образуют репликационную вилку, и каждая из них становится матрицей, на которой синтезируется новая комплементарная цепь. В результате образуются две новые двуспиральные молекулы ДНК, идентичные родительской молекуле.

В моей работе меня интересовало место начала репликации.

Начало репликации ДНК можно отождествить с местом, в котором начинается его раскручивание белками. Это место (сайт, site) называется ориджин (Origin, в дальнейшем **ORI**). Экспериментальными методами с различной точностью во многих ДНК найдены эти сайты, и они представляют собой отрезки нуклеотидов длиной от 10 до 300. Эти сайты являются особенными, так как репликация всегда начинается именно в них. Проблема заключается в том, что на данный момент наука не умеет вычислять местоположение ORI, как результат некой функции от ДНК.

Если бы мы научились это делать, то это стало бы прорывом не только в биологи, но и в медицине. Можно бы было попробовать предотвращать репликацию раковых клеток, что могло бы стать ключом к лечению смертельной болезни. Также можно было бы научиться регенерировать клетки и управлять ростом организмов.

Факторы

Мы с руководителем обозначили некоторые факторы, которые, по его мнению и мнению других исследователей в этой области, могли бы приблизительно определять местоположение ORI. Назовем **факторами** некоторые функции от конечной последовательности нуклеотидов, определяющие некую качественную характеристику последовательности.

Определять местоположение сайта ORI будем путем выявления некой зависимости между факторами на ORI-содержащих и ORI-несодержащих последовательностях.

Были предложены для начала 5 факторов:

1. CpG –острова
2. Факторы транскрипций
3. Палиндромы
4. Температура плавления
5. Геометрический фактор (кручение, кривизна)

Сейчас я распишу чуть подробнее о каждом факторе.

CpG-острова

В генетике , CpG островами обычно называют участок ДНК длиной примерно 200 пар нуклеотидов и выше, в котором кол-во GC-пар превышает 50%. Символ «p» относится к фосфорной связи между цитозином(C) и гуанином(G).

CpG острова характеризуются содержанием динуклеотидов CpG по крайней мере 60%, в то время как среднее по молекуле значение их содержания равно ~ 4-6%, а остальная часть генома содержит концентрацию CpG ~ 1%. Это явление называется CG подавлением.

CpG острова как правило, находятся внутри или вблизи сайтов начала транскрипции генов (процесс синтеза РНК).

Факторы транскрипции

Факторы транскрипции (транскрипционные факторы) — белки́, контролирующие перенос информации с молекулы ДНК в транскрипцию путем связывания со специфичными участками ДНК.

В моей работе эти факторы являют собой несколько специфичных коротких последовательностей нуклеотидов, вхождение которых в участок ДНК должно увеличивать вероятность появления ORI.

Палиндромы

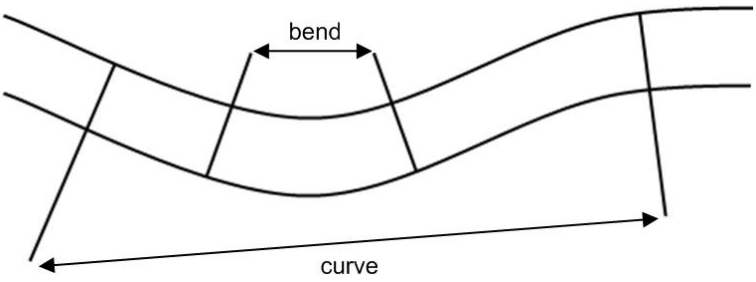
Это пары участков ДНК обратнокомплиментарных друг другу. Этот фактор может влиять на ORI из соображений, что между парой обратнокомплиментарных создаются между комплиментарными нуклеотидами и это может послужить поводом к началу репликации.

Температура плавления

Температура плавления это одна из основных характеристик молекулы ДНК - температура, при которой происходит диссоциация 50% двойной спирали, специфична для ДНК данного вида организмов, т.к. зависит от нуклеотидного состава и ее общих размеров. Она отражает AT/GC-соотношение в молекуле, т.к. пара G-C имеет 3 водородные связи (A-T - 2) и

взаимодействие между нуклеотидами этой пары более сильное, - соответственно, она положительно коррелирует с долей пары G-C в молекуле ДНК.

Геометрический фактор (кручение\кривизна)



Считается, что чем более изогнут, искривлен участок молекулы ДНК, тем легче белкам начать расплетать ее в этом месте.

Постановка задачи

Моя глобальная задача состояла в том, чтобы разобраться в проблеме нахождения ORI, сформулировать возможные варианты решения и реализовать их на практике.

Так как проблемная область очень широка и в ней практически нет проторенных дорожек, другой моей задачей являлось сформулировать для себя ряд мини-задач, которые бы помогли решить глобальную.

Для удобства дальнейшего исследования я поставил себе цель сделать программный инструмент, который бы позволил гибко и удобно получать последовательности нуклеотидов из общей базы в интернете и анализировать их. Для этого нужно было научиться программно вычислять все 5 факторов, и графически выводить результаты для анализа факторов на глаз.

На основе результатов вычисления факторов также требовалось реализовать некоторое машинное обучение для определения местоположения ORI с некоторой приемлемой вероятностью. Для этой цели я выбрал искусственную нейронную сеть.

Реализация

Для начала нужно было определиться как и откуда брать последовательности нуклеотидов. Мы решили брать их из геномного браузера от организации ENCODE, который находится по адресу http://genome.ucsc.edu/, размеры скачиваемых последовательностей менялись в процессе исследования (от 200 до 1000 нуклеотидов).

Также мы обладали списком известных ORI по версии статьи Genomic study of Replication Initiation in Human. Были выделены 16 наиболее ярко выраженных ORI для обучения. Отрицательные примеры брались либо сдвигом адреса известного ORI на несколько тысяч нуклеотидов, либо просто взятием случайного адреса в хромосоме, так как количество сайтов ORI и их размеры несоизмеримо малы по сравнению с размерами всей хромосомы.

Далее нужно было научиться вычислять удобно все факторы. В итоге для каждого фактора был написан Python-скрипт для его вычисления. Какие-то скрипты были написаны уже до меня: температура плавления, CpG-острова, факторы транскрипции – я их только изменил, чтобы факторы вычислялись с усредненными установками.

Температура плавления вычислялась следующим образом, «по окнам»: брался отрезок длиной 30 и, начиная с первого нуклеотида, с шагом 1 вычислялась средняя температура на отрезке.

CpG-острова вычислялись через отдельную программу, которая выводила результатом все острова и их CpG насыщенность.

Факторы транскрипции вычислялись просто поиском подстрок из определенного файла «Tf.txt» в строке, которую составляет наша последовательность.

Стояла задача научиться считать остальные два фактора, для этого требовалось найти инструмент, в котором процесс вычисление факторов можно было бы запустить программно. Было просмотрено множество инструментов (по большей части Web), которые умеют вычислять их, но мой выбор остановился на приложении eMboss, которое имеет как GUI, так и консольный интерфейс. Остальные инструменты либо имели только GUI интерфейс, либо не имели толковой документации как ими пользоваться в качестве сервисов.

Итак, с помощью eMboss были вычислены остальные факторы, и я приступил к написанию нейронной сети. Мной была выбрана библиотека Neurondotnet для ее реализации, которая хорошо себя зарекомендовала, в которой есть все нужное для обучения и тестирования сети, а также потому что .Net – привычная и удобная для меня платформа.

Я использовал простую трехслойную нейронную сеть, реализующую метод обратного распространения ошибки.

Первый подход для обучения сети был довольно «лобовой». Я попробовал скачать не очень длинные последовательности(около 500

нуклеотидов) и вычислял количественную характеристику всей последовательности по 5ти факторам. Это значит, что в качестве входных векторов в нейронную сеть я использовал следующие значения: количество CpG-островов, количество палиндромов, количество факторов транскрипции, среднее значение температуры в последовательности и среднее значение кручения \ кривизны.

Результат обучения был неутешительный. Сеть на последующих тестах практически не различала ORI-содержащие и ORI-несодержащие последовательности. Не было найдено никакой зависимости наличия ORI в последовательности от факторов.

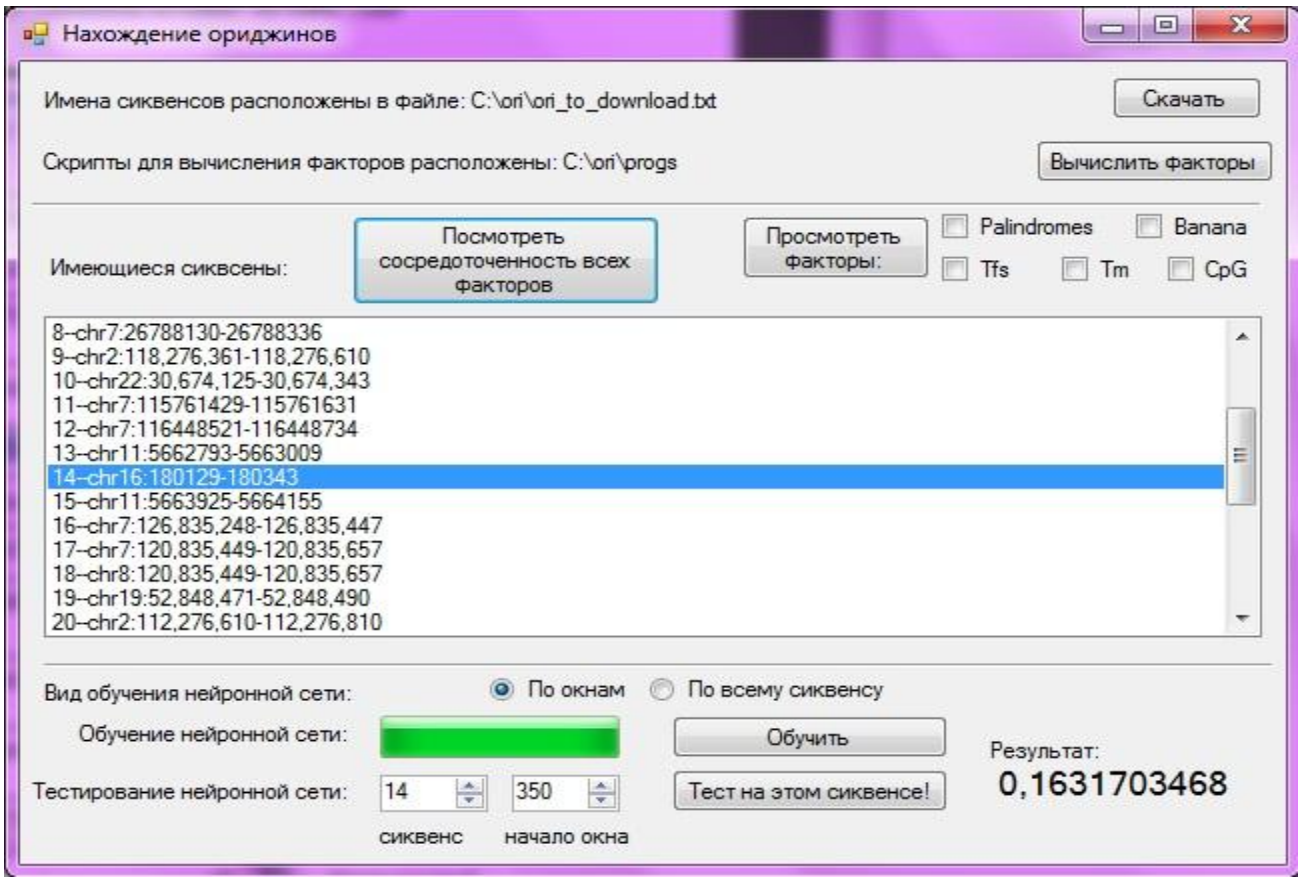
После этой неудачи было решено изменить логику подсчета факторов. Теперь я руководствовался не количественной характеристикой факторов, а локальной сосредоточенностью этих факторов в последовательности.

Факторы теперь вычислялись по окнам. В качестве функции от результатов «островных» факторов таких как CpG-острова, Палиндромы, Факторы транскрипций я использовал следующую величину: «Сумма 3х минимальных расстояний до ближайшего острова, притом что, если какое-то из 3х расстояний >50, то слагаемое заменяем на 50». Соответственно для обособленных участков, находящихся вдали от фактора было крайнее значение 150.

Температура плавления и так вычислялась по окнам, поэтому там искать функцию преобразования не пришлось. Геометрический фактор я взял средним значением в окне.

Обучение на сей раз опять ничего не дало.

После второй неудачи я написал график сосредоточенности всех факторов внутри одной последовательности.

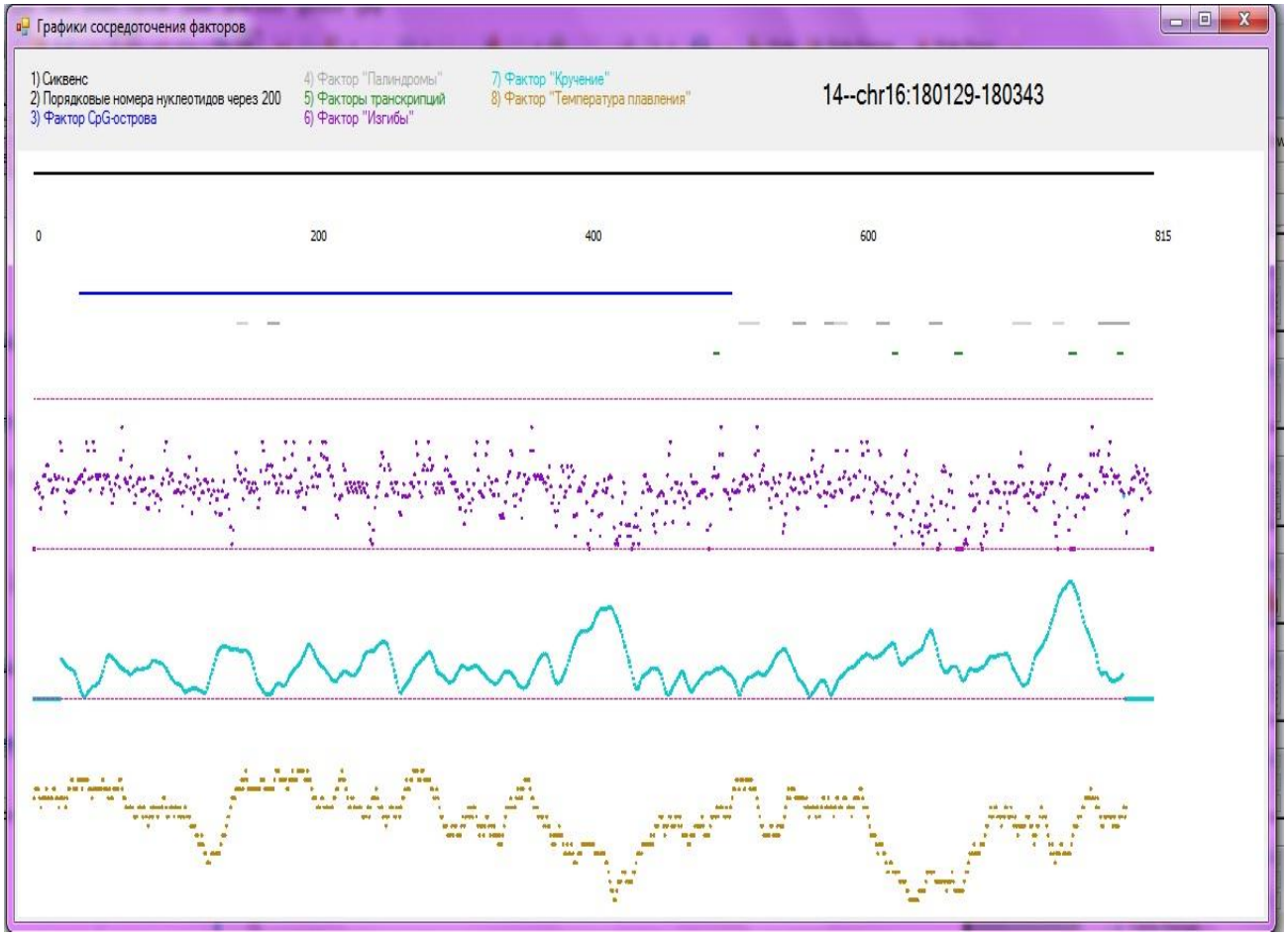


Результаты

В итоге я написал удобное приложение, которое позволяет скачивать последовательности из базы в интернете, вычислить все 5 факторов, просмотреть результаты вычисления факторов в текстовом варианте, а также для каждой последовательности в отдельности просмотреть график сосредоточенности факторов.

Также в рамках этого же приложения была реализована искусственная нейронная сеть, которая обучалась двумя способам (количественно на всю последовательность и локально сосредоточенно в окнах), но при этом не смогла выявить различия между ORI и не ORI.

По совокупности визуального анализа и результатов обучения сети был сделан довольно важный с точки зрения биологии вывод, что эти факторы не могут являться детекторами местоположения ORI.



Список литературы

1. N., Taylor C.M., Malhotra A. and Dutta A. Genomic Study of Replication Initiation in Human Chromosomes Reveals the Influence of Transcription Regulation and Chromatin Structure on Origin Selection // Mol Biol Cell. 2010. V. 21(3). P. 393-404.
2. http://neurondotnet.freehostia.com/manual/index.html
3. http://ru.wikipedia.org/wiki/Ori
4. http://ru.wikipedia.org/wiki/%D0%A2%D0%BE%D1%87%D0%BA%D0%B0_%D0%BD%D0%B0%D1%87%D0%B0%D0%BB%D0%B0_%D1%80%D0%B5%D0%BF%D0%BB%D0%B8%D0%BA%D0%B0%D1%86%D0%B8%D0%B8
5. http://en.wikipedia.org/wiki/CpG_island